Translating Low-Level Features into Student Friendly Explanations using XAI

Anonymous Author(s)

Abstract

An alarming decline in *Gaeilge* proficiency has renewed education efforts, fueled by the rise of AI in recent years. This research presents the development of a deep learning framework designed to evaluate spoken Gaeilge (henceforth *Irish*) proficiency of learners by quantifying student voice sample similarity to everyday Irish speaker truth samples, combined with explainable AI (XAI) techniques used to generate proof-of-concept student-friendly feedback. While automatic speech recognition (ASR) systems for high-resource languages like English have access to large, detailed datasets, the Irish language relies on a limited availability of speech resources, that challenges traditional model development. We address these gaps through a human-centered approach that employs a triplet-loss based Convolutional Neural Network (CNN) using spectrograms for similarity scoring, and XAI methods like using Mel-frequency cepstral coefficients (MFCCs) and Gradient-weighted Class Activation Mapping (Grad-CAM) to compare and contrast student attempts. Our evaluation combines technical metrics with visualisations, to make the learning process intuitive and simple for students. Results demonstrate that the model can successfully identify the similarity between student samples and truth samples, supplemented by power-transformed similarity scoring to help interpretability. This work contributes a scalable foundation for low-resource language education, demonstrating the potential for AI-assisted tools to aid in preserving languages like Irish while positioning XAI as a bridge between deep learning techniques and personalised, accessible learning for all students.

CCS Concepts

• **Computing methodologies** → **Neural networks**; **Modeling methodologies**; • **Security and privacy** → **Social aspects of security and privacy**.

Keywords

Irish Language, Explainable AI, Convolutional Neural Network, Speech Recognition, Educational AI, Triplet Loss

ACM Reference Format:

Anonymous Author(s). 2025. Translating Low-Level Features into Student Friendly Explanations using XAI. In *Proceedings of Human-Centered AI Education & Practice Conference (HCAI-ep '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HCAI-ep '26, Maynooth, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The evaluation of spoken language proficiency is inherently difficult, as the teachers would have to give feedback on individual learners' speech in order to encourage learning, leaving a majority of students without precise, actionable feedback. This challenge is particularly acute for minority languages like Irish, where in 2016, only 1.5% of the population were speaking Irish daily, and where limited resources compound learning difficulties [16] [18]. Many studies have focused on speech recognition and speech evaluation due to their importance in day-to-day life [5] [12]. While ASR systems have achieved remarkable success in high-resource languages, their performance remains poor for diverse low-resource contexts. This research aims to alleviate this problem by developing a solid CNN model, supported with XAI techniques, which can be used to help students understand how they have performed in comparison to expert-level speakers, and to visualise the aspects of their attempts that define their similarity score.

There are some existing projects in this space, such as the ABAIR initiative, which outlines three core aspects of speech technology for the Irish language: the provision of technology-oriented language resources; the thorough development of core speech technologies; and the development of technology applications, which exploit both the technologies and the linguistic resources [3]. Such platforms highlight the untapped potential of embedding-based voice analysis in low-resource languages. The system designed by ABAIR is unique to each student, by using spectrograms generated from the FastPitch acoustic model. The model predicts pitch contours during inference, which is crucial for natural-sounding speech, and remains flexible for different speakers [22]. The benefit of using spectrograms can be further strengthened by using parallel additional features like MFCCs, raw waveforms and embeddings.

Complementary features such as MFCCs further explain spectral outputs, capturing nuances of expression quality, nasal sound, and consonant-vowel transitions [19]. These techniques form the base of modern speech processing systems, yet their application requires careful preprocessing to ensure reliability. In this research, we followed necessary steps to ensure comprehensive feature extraction for speech data. We used the lossless audio format *.wav*, handled silent audio segments appropriately, and normalised sampling rates and volume. By following these steps, we ensured that the data was fully prepared and standardised across all phrases, ready to be used in MFCC and spectrogram generation, and model training.

While deep learning models excel at pattern and speech recognition, their "black-box" nature hinders educational use. Using a CNN, and cosine similarity between triplet-loss embeddings, we create a framework for similarity scoring between student and expert samples by comparing spectrograms and quantifying the similarity score. However, numerical scores alone provide insufficient guidance for improvement. This research integrates XAI to highlight

low-level acoustic features and provide learner-focused visualisations. Grad-CAM localises spectral regions influencing similarity judgments, revealing where pitch, duration, or spectral energy differ from expert pronunciations [1]. One study regarding ASR in the Irish language describes how: for a language with such distinctive, phonemic vowel length and dialectal variations, such granularity is a requirement [16]. This research establishes a pipeline for:

- Assessment: Quantifying pronunciation similarity via CNN performance, spectrograms, MFCCs, and similarity score distribution.
- Explainability: Using XAI to convert model outputs into visual and textual feedback. This includes Grad-CAM, power-transformed similarity score and MFCCs.
- Optimisation: Exploring different model depths, configurations and parameters to overcome data scarcity issues, training times and model re-evaluation.

By transforming speech spectrograms into explainable diagrams and visuals, this work advances both teaching tools for low-resource languages like Irish and interpretable AI methodologies. The following headings detail related works, dataset curation, methodology, results, ethics and limitations.

2 Related Works

This section describes advancements in speech technology for language learning, feature extraction, neural network architectures, and voice analysis applications, contextualising our research on triplet loss-based speaker verification for Irish, within the broader speech technology landscape.

As part of the ABAIR initiative, the *Fotheidil* system, developed as a web-based transcription system for the Irish language, describes some challenges and solutions presented by low-resource languages, particularly the lack of high quality audio material that can be used train an audio system in a supervised manner. They go on to discuss a possible remedy, in Noisy Student Teacher training (NST), where the model is used to generate pseudolabels for the unlabelled dataset, which is combined with the labelled data to create a new training set for the student model. The *Fotheidil* system also describes the use of MFCC features for their experiment, similar to our research [11].

MFCCs and log-mel spectrograms remain the standard for speech feature representation due to their ability to encode spectral characteristics, which are critical for speaker discrimination and speech related projects. Librosa-derived MFCCs form the base of traditional speech emotion recognition (SER) systems by capturing unique linguistic cues. Despite the advancements in deep neural networks (DNNs) and CNNs for speech recognition, noise robustness remains a significant challenge. One paper highlights that performance degradation occurs in DNN/CNN-based systems under acoustic mismatches, such as varying background noise, as evidenced by studies on datasets like Aurora 2 and TIMIT. This limitation underscores the need for improved noise-robust techniques, such as strategic data augmentation or hybrid model integration, to enhance system adaptability in diverse conditions [13].

Augmentation solutions include Gaussian noise injection, speed perturbations or pitch shifting, traditionally used to address cost and resource constraints, data imbalance, and data annotation problems. One study, that developed a CNN with insufficient data, showed that data augmentation effectively increased the model accuracy from 63% to 97%. Advancements in Low-Rank Adaptation (LoRA) techniques enable parameter fine-tuning of pretrained models by projecting weight updates into low-rank subspaces. This approach is particularly effective for adapting English models to under-resourced languages like Irish using minimal data. Recent enhancements demonstrate LoRA's success in audio applications, addressing dependencies and spectral characteristics while maintaining efficiency. Such techniques can assist in facilitating learning processes for students [4] [10].

The development of intelligent Computer-Assisted Language Learning (iCALL) platforms has not only transformed teaching approaches, but has also strengthened the research into the speech model development field with healthy, noise-conscious systems; particularly for low-resource languages. An Scéalaí ("The Storyteller") exemplifies this innovation, utilising speech synthesis and Natural Language Processing (NLP) to support Irish learners. This web-based system enables learners to write text, and receive synthetic voice feedback, and compare their recordings against these references; a process called aural proofing. Crucially, the platform archives text-recording pairs as snapshots, allowing learners to track progress over time. Empirical evaluations confirm its efficacy in developing acute awareness of a language's speech structure. [14] [15].

Building on these established approaches, speech composition is primarily learned through spectral representations that capture the essential acoustic patterns within short time windows. While CNNs excel at identifying local features from MFCCs and spectrograms, they can also learn other characteristics that span entire utterances, like a student's unique accent, speaking rhythm, or pronunciation habits across multiple words. One approach combines feature maps from different layers of the CNN, creating rich speech embeddings that capture everything from the fine-grained acoustic details that distinguish phonemes, to the higher-level patterns that characterise the user speech. By creating embeddings at the utterance level, information is gathered from many spectral frames together, smoothing out the noise and inconsistencies. This research examines what each layer contributes to these embeddings, investigating how acoustic information transforms as it flows through the network; knowledge that could prove valuable for developing XAI feedback systems that help students understand where their pronunciation differs from native, or otherwise expert speakers [17]

Voice analysis in language learning also extends to automatic pronunciation assessment (APA), where fine-grained feedback facilitates learner awareness and supports educational improvement for students. One paper [2] describes an automatic pronunciation scorer trained on a human-rated dataset, achieving strong correlations, nearly matching human reliability; making it promising for operational and teaching use. These advances underscore how APA systems, enriched by embeddings and multi-level scoring, offer precise and interpretable feedback, aligning with the goals of low-resource learner pronunciation development. An Scéalaí similarly leverages voice logging but focuses on phonological development rather than affective states. Our work bridges these domains by

proposing speaker embeddings as objective metrics for accent refinement, demonstrating how voice technology can concurrently serve teaching and diagnostic roles [20].

While these works demonstrate the pedagogical value of speech technology for the Irish language, they prioritise output correction over speaker-specific analysis. Our research integrates comprehensive pre-processing, and explainable, understandable feedback through saliency maps and visualisations.

3 Data

The dataset chosen for the study comes from a Gaeltacht school in the west of Ireland. The data consists of twelve phrase folders, each containing a sample folder (from students), a truth folder (from multiple expert speakers), and expert-identified bad samples. The audio files are contained within these folders, and these files are used to create the triplets that are used in model training.

Phrase	Truth	Student Sample	Bad Sample	Total
1	12	77	16	105
3	12	103	16	131
4	12	117	16	145
5	12	139	15	166
9	12	123	12	147
12	12	82	13	107
13	12	61	10	83
15	12	71	7	90
17	12	74	5	91
22	12	80	16	108
27	12	73	10	95
30	12	81	6	99

Table 1: File counts for each phrase

The audio files are a mixture of .mp3 and .wav files, ranging from around 1100ms to nearly 9000ms in duration, with a standard deviation between 600ms and 1000ms. The files in the folders are quite imbalanced, as there are only 12 truth files in each phrase folder, dwarfed by the files from students, ranging between 61 to 123 samples. The identifed "bad samples" number between 5 to 16 per phrase. Initial analysis shows us that there the audio files in the dataset were recorded on standard recording devices with sampling rates between 44kHz and 48kHz, and at a wide spectrum of different volume levels. Leading and trailing silences were also identified, at an average silence of 1244ms per file, and standard deviation of 661ms. These were removed to not only speed up the processing and modelling aspect, but also for standardisation, as discussed in the introduction.

4 Methodology

This research follows a systematic pipeline integrating data preprocessing, triplet generation, deep learning, and XAI to transform low-level acoustic features into understandable visualisations for students. The methodology is comprised of three phases:

4.1 Data Preprocessing

All files are converted to lossless .wav format, and leading and trailing silence is trimmed to increase overall efficiency and focus on the relevant part of the audio. We standardise the sampling rate to 16kHz, as one experiment discovered that 16kHz produced the lowest word error rates for Hidden Markov Models (HMMs), used for speech recognition under noisy conditions [8]. Volume normalisation was standardised at -20dB following established practices in deep learning for speech processing [9]. This preprocessing step removes variable acoustic conditions by standardising the statistical moments of the data, allowing the model to focus its learning capacity on more discriminative phonetic features rather than amplitude variations. Research indicates such normalisation accelerates model convergence while improving generalisation across recording conditions. Augmentation was explored, however, as triplet loss was used, we were able to generate nearly 2000 triplets per phrase using permutations, therefore nullifying the need for augmentation.

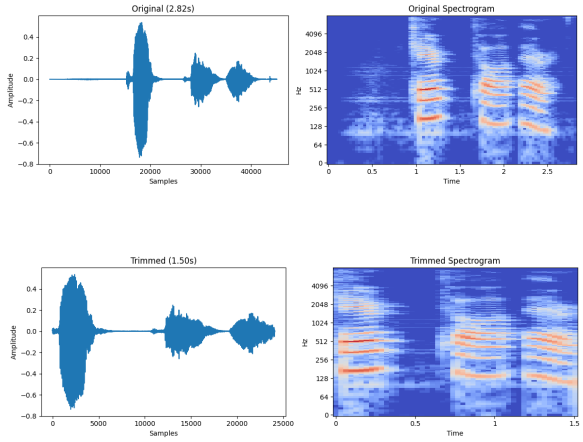

Figure 1: Before (above) & after (below) pre-processing; silence removal, volume normalisation & sampling rate standardisation.

4.2 Feature Extraction & Representation

Human loudness perception is logarithmic, on a decibel scale, and log-compression amplifies low-energy sounds. Therefore, we used the established standards in speech processing; A 25ms window size to balance time and frequency resolution. Hop length of 10ms provides a large overlap between frames, minimising information loss. Mel bins of 64 optimise granularity and efficiency. Fewer bins lose phonetic detail; more bins introduce redundancy. Studies show that 64 bins can effectively capture human-perceivable frequencies [6]. MFCCs and their deltas were captured to encode key vocal detections; like nasal sounds or the ending of a word.

4.3 Model Development

The core of our approach uses a CNN to quantify pronunciation similarity between student and expert recordings. The network processes triplets (Fig. 2) of log-mel spectrograms of the same Irish phrase through a shared CNN backbone, generating embedding

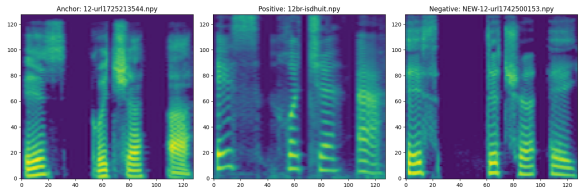

Figure 2: Sample triplet, used to train the model. Left to right: Anchor, Positive & Negative spectrogram

vectors for each input. These embeddings capture essential acoustic patterns, like vowel structure, spectral energy distributions, and prosody. As the model uses triplet loss, embedding distances are minimised for expert-expert pairs, while maximising separation for expert-student mismatches. During inference, cosine similarity between embeddings is converted to a 0-100% score, which can be understood as the % similarity the new sample is when compared to the expert sample. The true cosine similarity result between 0 and 1 is power transformed; for example, squared for samples closer to the truth, and cubed for poor samples. This helps the final similarity score become more intuitive for students, punishing poor results and strengthening good results.

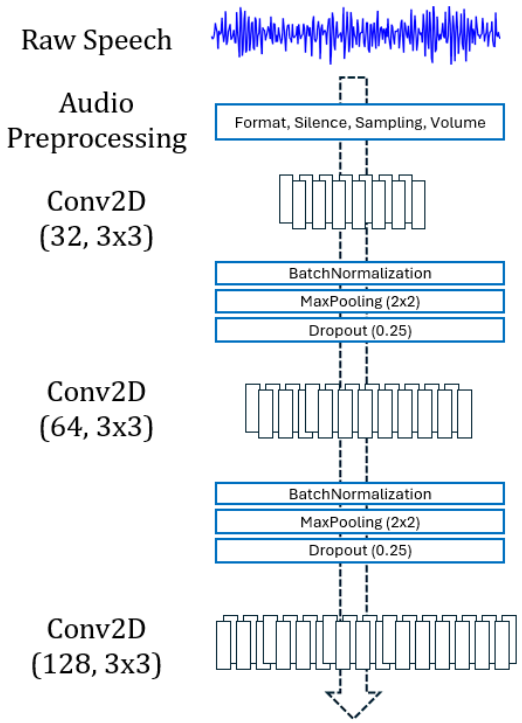

Figure 3: Convolutional layers of the model, the convolutional features are flattened into a vector, then passed through a dense 128-unit ReLU layer with dropout

We chose a CNN architecture (Fig. 3) with feature extraction that demonstrates a thoughtful approach to spectrogram analysis for pronunciation evaluation. The model's design reflects several key considerations for audio processing: the three convolutional blocks with increasing filter depth (32, 64, 128) systematically capture acoustic patterns, from spectral features to higher-level phonetic structures. The inclusion of batch normalisation after each convolutional layer normalises the activations of each layer, which is particularly valuable given the natural variability in speaker inputs, while the consistent dropout rate (0.25 in convolutional layers, 0.5 in dense layers) adds emphasis to regularisation, which is an essential consideration given the scarce dataset used. The max-pooling layers' 2x2 reduction strategically maintains spatial information while controlling complexity. Spatial attention was tested as an experimental approach to improving feature localisation within spectrograms, however it was deemed unsuitable due to input size issues and pipeline incompatibility. The final dense layer's 128-unit size creates an embedding space suitable for similarity comparisons while preserving discriminative features. This architecture effectively balances computational efficiency with model capacity, avoiding excessive parameters while remaining aware of the limited training data. The Rectified Linear Unit (ReLU) activations throughout promote efficient training while remaining robust towards complex audio patterns.

5 Experimental Results

We trained and evaluated four different models, ranging from a basic 2-layer architecture to a 4-layer architecture with dropout.

CNN-A is a straightforward baseline CNN with two convolutional layers and a fully connected layer of 128 units. It achieved a good sample similarity score of 95.09% and a bad sample similarity of 77.76%, with 14 false positives and 2 false negatives. The model effectively separated good and bad embeddings, as shown by a large difference between mean anchor-positive (61.50) and anchor-negative distances (772.55). However, the high standard deviations indicate inconsistency in the embedding space, especially for ambiguous samples, and this is reflected in its accuracy of 0.9207 ± 0.0063.

CNN-B builds on the baseline by adding depth and regularisation, using three convolutional blocks with batch normalisation and replacing flattening with global average pooling. This model improved similarity scores for both good (99.95%) and bad (98.3%) samples and reduced false positives from 14 to 10. However, false negatives increased sharply to 68, indicating a more conservative prediction behaviour and lower recall, which is clearly evidenced by its sensitivity score of just 0.5781. This trade-off resulted in the lowest overall accuracy among the four models at 0.7435 ± 0.0121. Embedding distances were much tighter and more consistent, with mean anchor-positive and anchor-negative distances dropping to 1.59 and 7.51, respectively, reflecting improved clustering quality despite the classification inconsistency.

CNN-C employs depthwise separable convolutions to reduce complexity and parameters. It achieved the best classification results with only 1 false negative and 4 false positives, outperforming both CNN-A and B. The statistical results confirm this superiority, with an excellent accuracy of 0.9289 ± 0.0074 and balanced sensitivity/specificity scores of 0.9128 and 0.9479 respectively. The embedding space was the most compact and well-separated, with

Model	Accuracy (Mean ± SD)	p-value	T-Statistic	Sensitivity	Specificity
CNN-A	0.9207 ± 0.0063	7.319E-171	-35.1423	0.8815	0.9583
CNN-B	0.7435 ± 0.0121	1.039E-50	-15.5456	0.5781	0.9115
CNN-C	0.9289 ± 0.0074	1.407E-87	-22.6332	0.9128	0.9479
CNN-Final	0.9954 ± 0.0016	1.626E-303	-59.2845	0.9948	0.9948

Table 2: Comparison of CNN models with accuracy, p-value, t-statistic, sensitivity and specificity.

mean anchor-positive distance at 0.154 and anchor-negative at 2.10, showing excellent clustering precision. Interestingly, the similarity scores for good and bad samples dropped significantly (to 57.72% and 15.39%), which in this context reflects the distinct embeddings, correlating with the low error rates.

Our final model's results demonstrate statistically significant and meaningful performance, validated on unseen data. The t-statistic of -59.28, near-zero p-value (1.626E-303), and perfect balance between sensitivity and specificity (both 0.9948) provide evidence that the model learned highly discriminative features of audio. Statistical significance is further evidenced by the extremely small standard deviation in accuracy (± 0.0016), indicating consistent performance across evaluations. These results confirm CNN-Final's ability to consistently learn meaningful distinctions relevant to speech data. Performance metrics; accuracy, sensitivity, specificity, precision, and F1-score are consistently above 95% across phrases. The Receiver-operating characteristic (ROC) and Area under the Curve (AUC) near 1 and high Matthews Correlation Coefficients further confirm the model's reliable classification ability. This balanced performance underscores the end goal of the research: to facilitate trustworthy feedback for Irish learners, while minimising false positives and negatives.

In addition to testing the model on data present within the original dataset, we also evaluated it's performance using external sources, each of which, provided one sample of each phrase. The aim of this test is to check how the model performs when exposed to new data. We compared the performance on five files per phrase; 5 seconds of silence, the new, external data files (denoted as T1, T2 and T3), and a random truth sample from the tested phrase. We ran each test 3 times due to the random selection of the expert file, and have taken the average of all runs. The silence and external files produced the same result three times. All similarity scores in table 3 are as a percentage, power-transformed with a threshold of 0.5 based on raw similarity. The results justify the per-phrase adjustment of the threshold for power-scoring in future work.

As we can see from the table above, the model is able to identify the difference between complete silence and external, new data on all three occasions. The similarity scores are consistently higher, with the highest being the expert file, as expected. In conclusion, the model's combined statistical strength, practical effect size, explainability, and generalisability mark it as a valuable tool for improving Irish language learning by helping students identify and correct poor speech patterns efficiently and reliably.

Phrase	Silence	T1	T2	T3	Expert
1	53.6	77.8	75.3	75.1	85.6
3	63.7	72.1	79.5	71.4	87.6
4	77.9	79.6	89.9	85.1	88.2
5	56.6	61.9	58.4	57.6	92
9	70.6	79.8	82.5	71.9	69.1
12	74.4	83.3	92.2	95.5	93.1
13	68.3	85.2	96.5	91.1	95.8
15	76.8	95.2	92.8	88.2	89.6
17	53	58	69.7	67.2	86.7
22	50.8	62.8	65.8	69.6	79.5
27	62	94.9	75.8	81.2	82.1
30	60.8	94.4	94	98.4	96.7

Table 3: Raw similarity scores (%) for silence, external and truth files

6 Explainability

Explainability is paramount in this study to ensure that outcomes derived from anonymised audio data are interpretable and actionable for teaching. Under the EU AI Act, educational tools involving minors are classified as high-risk systems (Art. 6), enforcing "human-understandable" explanations for outputs. To comply with this regulatory requirement, model transparency was prioritised using three different techniques.

GradCAM saliency maps aim to clarify "black-box" decisions by highlighting critical regions influencing predictions (Fig. 4), such as mispronunciations. This aligns with the AI Act's emphasis on technical documentation for high-risk systems (Art. 13).

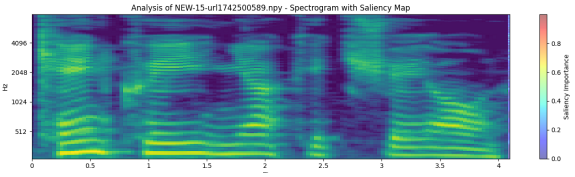

Figure 4: Sample saliency map from phrase 13.

A student will record and send an attempt through the preprocessing pipeline to be evaluated against the model, and returned is a similarity score, GradCAM saliency map and a MFCC comparison (Fig. 5). It is important to detail the negative parts, such as the fact that the model is not used in production, or that the model is not trained on diverse data (i.e. augmented or synthetic).

On the positive side, MFCCs enable standardised spectral comparisons of phonetic features, paired with saliency maps to illustrate model decisions. This dual approach supports the AI Act's transparency obligations and the UN's Sustainable Development Goal (SDG) 4, of inclusive and equitable quality education for all.

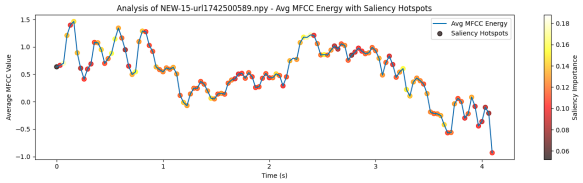

Figure 5: Sample MFCC with "hotspots", annotated where saliency and MFCC cross.

Collectively, these methods bridge technical complexity with stakeholder accessibility, enabling teachers to validate results and refine strategies, supporting the AI Act's Article 14 (human oversight), and SDG 4.7's call for education while promoting sustainable development [21] [7].

Ethical considerations in this study combine data governance with educational transparency. The anonymisation of the data and General Data Protection Regulation (GDPR) compliant handling reflect core principles of the EU AI Act, which emphasises data minimisation and privacy-by-design in AI deployments. Furthermore, the combined approach, where teachers and students alike engaged with audio sourcing, supports SDG 10 (Reduced Inequalities) by involving all stakeholders. Crucially, the explainability tools served an ethical function beyond transparency: GradCAM and MFCC methods enabled equitable assessment across a diverse group of speakers, although, it was not tested fairly (see Section 7). These points align with the EU AI Act's prohibition of "exploitative systems", and aim to provide an accessible platform for all. [7] [21].

7 Limitations

Several limitations deserve acknowledgment and explanation in this research. First, while an attention mechanism was initially explored to enhance model interpretability, technical integration challenges arose between the attention layer's input and the established pipeline, thus limiting the full exploration of training methods. Future work could explore the different model architecture available, including an attention layer. As a result, grad-CAM saliency maps were prioritised as a more stable alternative for highlighting critical regions within the speech data, though this forfeited the attention layer's advantage in mapping feature-level dependencies across individual samples.

Second, the absence of audio augmentation (like pitch shifting or noise injection) restricted the range of diverse data that could have been used to create a more robust model, potentially limiting model awareness to natural variations in acoustics, e.g. in a classroom, or unique student vocal characteristics. This decision stemmed from ethical and technical caution: augmenting anonymised audio data risked distorting specific speech traits fundamental to the study's goals.

Third, the limited amount of dialect-specific training data narrowed the model's exposure to speech pattern and general audio diversity, potentially amplifying bias toward dominant speech patterns. As identified in section 3, the dataset is limited one Gaeltacht region, and the statistics show that the files needed comprehensive preprocessing, which excluded augmentation.

Additionally, the model was not tested on students, and as such, the results cannot be verified, or considered a true reflection of model performance with new data. The testers who provided the test data were not primary-level students, but much older and somewhat familiar with the phonetics and pronunciation style of the Irish language. Furthermore, the model was not put into production and could not be used to test a wide variety of samples.

Collectively, these limitations highlight trade-offs between technical ambition, ethical safeguards, and reality, directing future work toward integrated attention architectures, ethically grounded data, synthetic or hybrid, and human-centered development and testing, by using correct and suitable testing techniques.

8 Conclusion

This study successfully developed and validated an explainable AI framework to transform anonymised Irish speech data into teachable, understandable insights for language education. Our final model achieved a mean accuracy of 99.54% (± 0.16%) with balanced sensitivity and specificity (0.9948). Although the results appear near-perfect, it it more probable that the model is overfitting and warrants additional fine-tuning in future work. By explaining model decisions through visual and numerical feedback, these methods enable students to conduct meaningful self-assessment. Ethically, the research strictly adhered to GDPR and the EU AI Act through comprehensive anonymisation while training, and adherence to EU and worldwide directives like the UN SDGs, by centering the needs of minority-language learners. If the model was put into production, the samples, student and expert, would not be played back to students, but rather based solely on similarity score, saliency maps and MFCCs in order to protect the source anonymity of the data.

An honest assessment must acknowledge the trade-offs made to prioritise data integrity over technical complexity. Limitations like not implementing attention mechanisms, the clear data scarcity, and the omission of audio augmentation were conscious choices that, while potentially limiting model performance, ensured a transparent and ethically-grounded research. These constraints open new avenues for future work: developing hybrid attention-saliency architectures for finer-grained feature mapping, pursuing ethically-sourced synthetic Irish speech data to overcome data scarcity, and implementing human-in-the-loop augmentation. While this work is technically sound, and explains *how* the model works and how it's results were generated, but future work could tackle the *why* at a much deeper level (i.e., pitch, nasality, accents). This project proves that AI can be effectively anchored in an educational, minority-language setting. The call to action is clear: researchers should continue to support such initiatives, investing in technology that honours heritage while complying with ethical standards.

References

[1] Adrien Bennetot, Ivan Donadello, Aya El Qadi, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Saranti, Silvia Tulli, Maria Trocan, Raja Chatila, Andreas Holzinger, Artur d'Avila Garcez, and Natalia Díaz-Rodríguez. 2021. A Practical guide on Explainable AI Techniques applied on Biomedical use case applications. (2021). http://arxiv.org/abs/2111.14260

[2] Danwei Cai, Ben Naismith, Maria Kostromitina, Zhongwei Teng, Kevin P Yancey, and Geoffrey T LaFlair. 2025. Developing an Automatic Pronunciation Scorer: Aligning Speech Evaluation Models and Applied Linguistics Constructs. *Language Learning* (2025).

[3] Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl. 2017. The ABAIR Initiative: Bringing Spoken Irish into the Digital Space. In *Interspeech 2017*. 2113–2117. doi:10.21437/Interspeech.2017-1407

[4] Hung-Chi Chu, Young-Lin Zhang, and Hao-Chu Chiang. 2023. A CNN sound classification mechanism using data augmentation. *Sensors* 23, 15 (2023), 6972.

[5] Nivja H De Jong. 2023. Assessing second language speaking proficiency. *Annual Review of Linguistics* 9, 1 (2023), 541–560.

[6] Gao Dongrui, Xue Tang, Manqing Wan, Guo Huang, and YONGQING ZHANG. 2023. EEG driving fatigue detection based on log-Mel spectrogram and convolutional recurrent neural networks. *Frontiers in Neuroscience* 17 (03 2023), 1136609. doi:10.3389/fnins.2023.1136609

[7] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. OJ L, 155 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689 Entry into force: 1 August 2024.

[8] Hans-Günter Hirsch and David Pearce. 2001. The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. In *ISCA Tutorial and Research Workshop (ITRW) on Automatic Speech Recognition: Challenges for the New Millenium*.

[9] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 2012. *Efficient BackProp*. Springer Berlin Heidelberg, Berlin, Heidelberg, 9–48. doi:10.1007/978-3-642-35289-8_3

[10] Jizhong Liu, Gang Li, Junbo Zhang, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. 2024. Enhancing automated audio captioning via large language models with optimized audio encoding. *arXiv preprint arXiv:2406.13275* (2024).

[11] Liam Lonergan, Ibon Saratxaga, John Sloan, Oscar Maharog, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2024. Fotheidil: an Automatic Transcription System for the Irish Language. *arXiv preprint arXiv:2501.00509* (2024).

[12] Sureshkumar Natarajan, Syed Abdul Rahman Al-Haddad, Faisul Arif Ahmad, Raja Kamil, Mohd Khair Hassan, Syaril Azrad, June Francis Macleans, Sadiq H Abdulhussain, Basheera M Mahmmod, Nurbek Saparkhojayev, et al. 2025. Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Engineering Journal* 16, 7 (2025), 103405.

[13] Sureshkumar Natarajan, Syed Abdul Rahman Al-Haddad, Faisul Arif Ahmad, Raja Kamil, Mohd Khair Hassan, Syaril Azrad, June Francis Macleans, Sadiq H. Abdulhussain, Basheera M. Mahmmod, Nurbek Saparkhojayev, and Aigul Dauitbayeva. 2025. Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Engineering Journal* 16, 7 (2025), 103405. doi:10.1016/j.asej.2025.103405

[14] Neasa Ní Chiaráin, Oisín Nolan, Madeleine Comtois, Neimhin Robinson Gunning, Harald Berthelsen, and Ailbhe Ní Chasaide. 2022. Using Speech and NLP Resources to Build an iCALL Platform for a Minority Language. In *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*. 109–118. doi:10.18653/v1/2022.computel-1.14

[15] Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2018. An Scéalaí: Synthetic Voices for Autonomous Learning. In *EUROCALL 2018 Conference Proceedings*. 230–235. https://eric.ed.gov/?id=ED590705

[16] Mengjie Qian, Hanne Berthelsen, Liam Lonergan, Andrew Murphy, Ciarán O'Neill, Neasa Ní Chiaráin, Chloé Gob, and Ailbhe Ní Chasaide. 2022. Automatic Speech Recognition for Irish: Testing lexicons and language models. In *2022 33rd Irish Signals and Systems Conference (ISSC)*. doi:10.1109/ISSC55427.2022.9826201

[17] Joanna Rownicka, Peter Bell, and Steve Renals. 2018. Analyzing deep CNN-based utterance embeddings for acoustic model adaptation. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 235–241.

[18] Joey Sanders. 2023. The Decline of the Irish Language in the Nineteenth Century. *The General: Brock University Undergraduate Journal of History* 8 (2023), 193–208.

[19] Saska Tirronen, Sudarsana Kadiri, and Paavo Alku. 2022. The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection. *Journal of Voice: Official Journal of the Voice Foundation* (2022). doi:10.1016/j.jvoice.2022.03.021

[20] Shinichi Tokuno. 2021. How Voice Analysis Technology Can Track the Psychological Impact of Covid. *CrisisNow* (2021). https://talk.crisisnow.com/how-voice-analysis-technology-can-track-the-psychological-impact-of-covid/

[21] Nations United. 2015. *Transforming our world: the 2030 Agenda for Sustainable Development*. Technical Report. New York. https://sustainabledevelopment.un.org/topics

[22] Adrian Łańcucki. 2021. FastPitch: Parallel Text-to-speech with Pitch Prediction. arXiv:2006.06873 [eess.AS] https://arxiv.org/abs/2006.06873